ISSN:0975-9646

An Analysis on Web Usage Mining For Internet Users

Priyanka Sharma[#1], Sumayya Khan[*2], Shilpa Singh[#3],Pooja Tiwari[#4]

[#]*Computer Engineering Department, Mumbai University*
Theem College of Engineering
Boisar(E), Palghar-401501

Abstract-Web Usage Mining is a branch of web mining. The data assembled from web access results in awfully large information represented in binary form. The data is grouped using divisive clustering method. The divisive analysis is one of the types of hierarchical method of clustering , the divisive analysis is used to separate each dataset from the clustered dataset. Here the algorithms D-Apriori and DFP are proposed to find the frequently accessed webpage from web log database and they will be comparatively analysed for the implementation of web usage mining and signifying which algorithm is more efficient than the other in terms of computational and scanning time.

Keywords- Apriori, Clustering, D-Apriori, DFP Algorithm, FP Algorithm and Web Usage Mining.

1. INTRODUCTION

Data mining is a technique used to deduce useful and relevant information to guide professional decisions and other scientific research. It is a cost-effective way of analyzing large amounts of data, especially when a human could not analyze such datasets.

Massification of the use the internet has made automatic knowledge extraction from Web log files a necessity. Information providers are interested in techniques that could learn Web users' information needs and preferences. This can improve the effectiveness of their Web sites by adapting the information structure of the sites to the users' behavior.

Recently, the advent of data mining techniques for discovering usage pattern from Web data (Web Usage Mining) indicates that these techniques can be a viable alternative to traditional decision making tools. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data and is targeted towards applications.

This project explores the use of Web Usage Mining techniques to analyze Web log records. We have identified several web access pattern by applying well known data mining techniques (D-Apriori and DFP Algorithm) to the access logs. This includes descriptive statistic and Association Rules for the portal including support and confidence to represent the Web usage and user behavior for that portal. The results and findings of this experimental analysis can be used by the Web administration in order to plan the upgrading and enhancement of the portal presentation. This project comparatively analyses both the data mining techniques namely D-Apriori and DFP algorithms.

2. HIERARCHIAL CLUSTERING

Hierarchical clustering is a process of cluster analysis which seeks to assemble a hierarchy of clusters. Strategies for hierarchical clustering are of two types namely Agglomerative analysis and divisive analysis.

In this study we use the Divisive analysis method to perform clustering of the web log file.

2.1 Divisive Analysis

This is a "top down" approach. All explanation start in one cluster and splits are performed recursively as one move down the hierarchy. Here the datasets are clustered using divisive analysis, the clustered datasets are split into a single cluster.

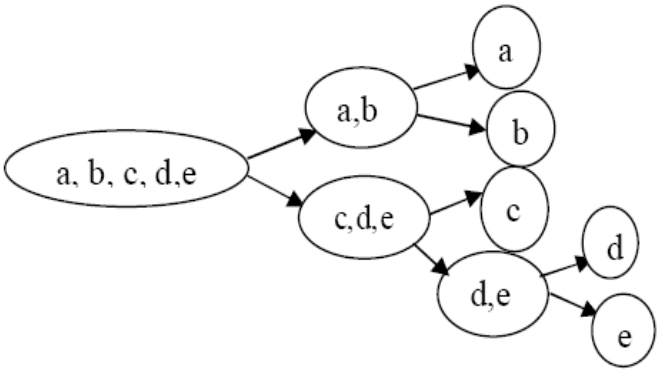

Figure 1: Divisive Analysis

3. D-APRIORI ALGORITHM

D-APRIORI stands for Divisive APRIORI algorithm. The output of the divisive analysis is given to the APRIORI algorithm, The algorithm attempts to find subsets which are common to at least a minimum number C (the cut off, or Confidence threshold) of the item sets. The system operates in the following three modules.

- Pre-processing module
- Apriori or FP Growth Algorithm Module
- Association Rule Generation
- Results

The pre-processing module converts the log file, which normally is in ASCII format, into a database like format,

which can be processed by the Apriori algorithm.

The second module is performed in two steps.

- Frequent Item set generation
- Rules derivation

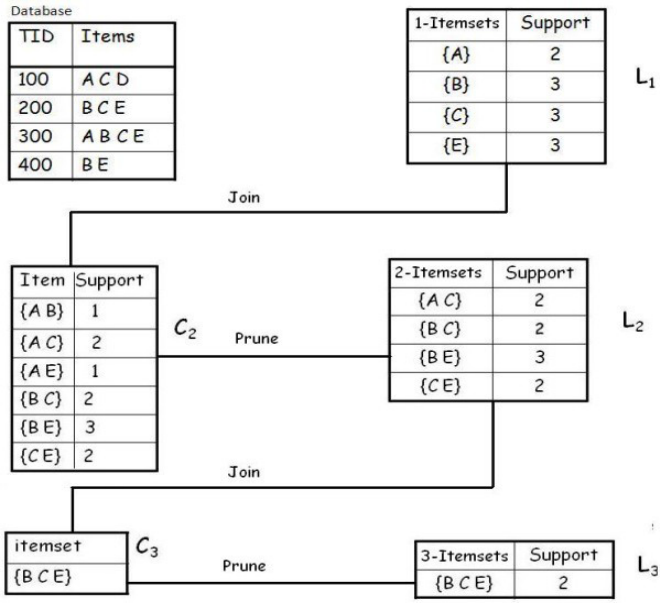

Figure 2: APRIORI Example

3.1 Advantages and Disadvantages APRIORI algorithm

Advantages:-

- Uses large Item set property
- Easy to implement
- Easily parallelized

Disadvantages:-

- It is costly to handle large number of item sets
- It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns.

4. DFP ALGORITHM

The DFP algorithm stands for Divisive FP growth algorithm. The FP Growth algorithm operates in the following four modules.

- Pre-processing module
- FP Tree an FP Growth Module
- Association Rule Generation
- Results

The pre-processing modules convert the log file, which normally is in ASCII format, into a database like format, which can be processed by the FP Growth algorithm.

The 2nd module is performed in two steps.

- FP Tree generation
- Applying FP Growth to generate association rules

FP tree is a compact data structure that stores important, crucial and quantitative information about frequent patterns.

The main components of FP tree are:

- It consists of one root labelled as "root", a set of item prefix sub-trees as the children of the root, and a frequent-item header table.
- Each node in the item prefix sub-tree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP tree carrying the same item-name, or null if there is none.
- Each entry in the frequent-item header table consists of two fields, (1) item-name and (2) head of node link, which points to the first node in the FP-tree carrying the item-name.

Second, an FP-tree-based pattern-fragment growth mining method is developed, which starts from a frequent length-1 pattern (as an initial suffix pattern), examines only its conditional-pattern base (a "sub-database" which consists of the set of frequent items co-occurring with the suffix pattern), constructs its (conditional) FP-tree, and performs mining recursively with such a tree. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree.

Since the frequent item set in any transaction is always encoded in the corresponding path of the frequent-pattern trees, pattern growth ensures the completeness of the result.

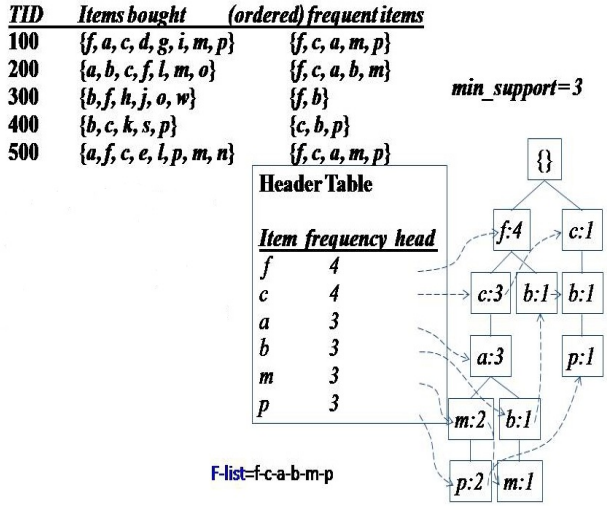

Figure 3: FP Tree generation

4.1 Advantages and Disadvantages of FP-GROWTH Algorithm

Advantages:-

- Uses compact data structure
- Eliminates repeated database scan
- FP-growth is an order of magnitude faster than other association mining algorithms and is also faster than tree-Researching

Disadvantages:-

- The main drawback of FP-growth algorithm is the explosive quantity of lacks a good candidate generation method.

5. RESULTS

The main objective of any system is the generation of reports. It has various uses. Some of them are,

- For the users, reports provide source of information required.
- They provide permanent hard copy of the results of transactions.

Careful consideration is being given in the designing of the reports as it helps in decision-making process. In the present work, the performance of the system is judged using two metrics. The first one is the amount of memory used and the second one is the time taken for the algorithm to create the association rules.

CONCLUSION

In this work the D-Apriori and DFP is proposed to analyse Web log records. The D-Apriori and DFP will provide useful information such as the user's browser behaviour and that can be used by the web administrator to incorporate content that is looked up by a number of users.
In our study, the results of both the algorithms will be comparatively analysed to see which of the two algorithms is more efficient.

REFERENCES

[1] Kotsiantis S, Kanellopoulos D., *Association RulesMining: A Recent Overview, GESTS InternationalTransactions on Computer Science and Engineering*,Vol.32 (1), 2006, pp. 71-82

[2] Agrawal R, Srikant R., "*Fast Algorithms for MiningAssociation Rules*", VLDB. Sep 12-15 1994, Chile,487-99, pdf, ISBN 1-55860-153-8.

[3] Mannila H, Toivonen H, Verkamo A I., "*Efficientalgorithms for discovering association rules." AAAI Workshop on Knowledge Discovery in Databases*(SIGKDD). July 1994, Seattle, 181-92.

[4] Tan, P. N., M. Steinbach, V. Kumar, *.Introduction toData Mining"*, Addison-Wesley, 2005, 769pp.

[5] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementation*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2005.

[6] P. Becuzzi, M. Coppola, and M. Vanneschi, *.Mining of Association Rules in Very Large Databases: A Structured Parallel Approach,"* Proc. Europar-99, vol. 1685, pp. 1441-1450, Aug. 1999.

[7] R. Jin and G. Agrawal, *"An Efficient Implementation of Apriori Association Mining on Cluster of SMPs,"* Proc. Workshop High Performance Data Mining (IPDPS 2001), Apr. 2001.

[8] J. Han and M. Kamber, *"Data Mining: Concepts and Techniques"* .Morgan Kaufmann Publishers, 2000.

[9] E.-H. Han, G. Karypis, and V. Kumar, *"Scalable Parallel Data mining for Association Rules,"* Proc. ACM SIGMOD 1997, May 1997.

[10] E-H. Han, G. Karypis, and V. Kumar, *"Scalable Parallel Data mining for Association Rules,"* IEEE Trans. Data and Knowledge Eng.,vol. 12, no. 3, May/June 2000.

[11] H. Cokrowijoyo, D. Taniar, *A framework for mining association rules in Data Warehouses*, Proc. IDEAL 2004, Lecture Notes in Computer Science, vol. 3177, Springer, Berlin, 2004, pp. 159.165.

[12] L. Dehaspe, L. Raedt, *Mining association rules in multiple relations*, Proc. ILP.97, Lecture Notes in Computer Science, vol. 1297, Springer, Berlin, 1997,pp. 125.132.

[13] Jiawei Han and Michelin Kamber, "Data mining Concepts and Techniques", Elsevier publication, Edition 2006.

[14] Rajan Chattamvelli, "Data Mining Methods", Narosa publications, Edition 2009.

[15] Santhosh Kumar and Rukumani,"web usage mining", ijana publication, vol.1, pages 400-404, Edition 2010.

[16] Ashok Kumar D, Loraine Charlet Annie M.C, " Web log mining using K-Apriori algorithm", ijca publication, vol.41 Edition march 2012.

[17] Shyam Sundar Meena, "Efficient discovery of frequent pattern using KFP-Tree from web logs", ijca publication, vol.49, Edition July 2012.

[18] G.Sudamathy and C.Jothi venkateshwaran, "An efficient hierarchical frequent pattern analysis approach for web usage", ijca publication, vol.43, Edition 2012.

[19] Jianhan Zhu, Jun Hong and John G. Hughes, "Page clustering: Mining conceptual link hierarchical from web log files for adaptive websites navigation", ACM publication, vol.4, Edition 2004.